Sara Santoriello

Il Meta-diritto dell'oversight board

(doi: 10.53227/105078)

Rivista di Digital Politics (ISSN 2785-0072)
Fascicolo 1-2, gennaio-agosto 2022

Ente di afferenza:
()

Sara Santoriello

Il Meta-diritto dell'*oversight board*

THE META-LAW ISSUED BY THE OVERSIGHT BOARD

Digital platforms experience new levels of complexity as their reach and interactions increase. Every day, thousands of pieces of content are classified as unsuitable for permanence on the web. The balance between forms of control and the actions of users has led Meta's administration toward the implementation of the Oversight board (Mob), a guarantor of freedom of expression and online safety that can provide independent judgment on compliance with the Community standards. Based upon the guidance given by the regulations, content moderation employs two approaches: horizontal, with user reports; vertical, through automated detection driven by Artificial intelligence and the supervision of reviewers, who are employed by the company. While content moderator is just one of the temporary jobs in the gig economy, Ai guarantees efficient control, identifying content classifiable as spam, but it risks encroaching on the field of decision making in certain cases requiring protection such as freedom of expression. Therefore, the Oversight board has a decisive role in determining «what to remove, what to leave, and why», although it does not directly affect the algorithms and market business strategies. Beginning with decisions published from 2020 to 2022, this article profiles the risks citizens face in exercising their rights in relation to the code and the predictive ability of algorithms to implement qualitative discriminations.

KEYWORDS *Meta Oversight Board, Administration, Social Media, Digital Platforms, Algorithms, Meta Ai.*

1. Introduzione

Uno dei principali problemi che sta sperimentando la società dell'informazione è un esponenziale e incessante lavoro di gestione dell'abbondanza (Franchi 2021, 57) dovuto alla composizione delle piattaforme. Gli algoritmi che sorreggono i social media sono stati progettati per adattarsi agli scenari mutevoli della frequenza di pubblicazione che non segue alcun criterio di linea editoriale stabilito a priori o rinegoziato. Pochi anni dopo la sua fondazione,

Sara Concetta Santoriello, Dipartimento di Scienze sociali – Università di Napoli, Federico II – Vico Monte della Pietà, 1 – 80138 Napoli, email: saraconcetta.santoriello@unina.it, orcid: 0000-0001-5313-8458.

ISSN 2785-0072

Facebook (ora Meta) introdusse una funzione che donava ai suoi utenti il potere di segnalare i contenuti non idonei alla circolazione nel *newsfeed*. Lo strumento della segnalazione diventava un campanello d'allarme sulla violazione delle norme in grado di portare all'attenzione dei suoi dipendenti la presenza di comportamenti perturbanti e accogliere il lavoro volontario di supervisione offerto dai cittadini. Con il tempo l'azienda ha potenziato le capacità del *machine learning* di monitorare e riconoscere i contenuti considerati pericolosi per la comunità digitale. La decisione finale sul destino dei contenuti in Rete, l'interpretazione delle norme e l'analisi dei casi sono state prerogative dell'azienda fino al 2020.

L'articolo analizza la composizione degli Standard della community, le regole stabilite dall'amministrazione per la corretta circolazione dei post e dei profili sulle sue piattaforme e come queste ultime agiscano qualora l'Ia o i moderatori dei contenuti commettano degli errori di valutazione. Nel caso in questione, è prevista una Procedura di ricorso che concede all'utente l'appello all'Oversight board, il Comitato per il controllo che funge da organismo indipendente in grado di poter dirimere le controversie ritenute rilevanti. Dal momento che nuove forme costituzionali emergono in momenti di debolezza e delegittimazione delle strutture politiche, sociali ed economiche pre-esistenti (Douek 2019), l'argomento rappresenta un'occasione per approfondire questioni aperte come autonomia e controllo dei sistemi di decisione autonomizzata e, in particolare, il potere creativo dei proprietari delle piattaforme di provvedere *ad hoc* alla propria giurisdizione. Questo precedente rende Mark Zuckerberg il fondatore di un «Meta-diritto».

2. Gli standard della community

I social media devono il proprio funzionamento alle interazioni tra gli utenti. Oltre due miliardi di persone hanno scelto di usufruire dei servizi messi a disposizione da Meta per raggiungere nuove cerchie e creare comunità, esprimersi liberamente e condividere esperienze. Le piattaforme che hanno sede negli Stati uniti possono scegliere i modelli di governance che ritengono opportuni per moderare (vale a dire come vagliare, classificare, filtrare e bloccare) i contenuti diffusi sul sito senza i vincoli del Primo emendamento[1] e senza es-

[1] Con cui la Costituzione degli Stati uniti sancisce il principio di terzietà della legge rispetto al culto e al suo libero esercizio, nonché la libertà di parola e di stampa e il diritto di appellarsi al governo: «Il Congresso non promulgherà leggi per il riconoscimento ufficiale di una religione, o che ne proibiscano la libera professione; o che limitino la libertà di

sere considerati civilmente responsabili come editori, secondo quanto stabilito dalla sezione 230 del *Communications decency act* del 1996 (Klonick 2020).

Al momento dell'iscrizione, l'utente accetta di rispettare le norme inserite all'interno degli Standard della community, che regolamentano la convivenza *inter pares* descrivendo cosa è consentito e cosa è vietato fare sulla piattaforma, con l'obiettivo implicito di dirimere le controversie. Come si legge nella sezione dedicata, «questi standard si basano sui feedback ricevuti dalle persone e sui consigli di esperti in ambiti quali tecnologia, sicurezza pubblica e Diritti umani» (*ibidem*) e forniscono la dovuta attenzione alla fattispecie per cui materiali ritenuti rilevanti e di pubblico interesse possano corrispondere a una deroga delle disposizioni generali. Le norme internazionali di riferimento sono menzionate nel Codice di condotta, dove si asserisce che: «Facebook si impegna a rispettare tutti i diritti umani riconosciuti a livello globale, che comprendono il diritto alla privacy, la libertà di espressione e tutti gli altri diritti indicati nella Carta internazionale dei diritti umani, nella Dichiarazione sui principi e i diritti fondamentali nel lavoro dell'Organizzazione internazionale del lavoro (Oil)».

Tenuto conto che internet, in quanto strumento, può catalizzare usi impropri mediante l'espressione, sono previste alcune limitazioni della libertà al fine di tutelare almeno uno dei seguenti valori: 1) autenticità; 2) sicurezza; 3) privacy; 4) dignità. A ciascuno vengono, inoltre, associati i rispettivi rischi: *fake identity*, intimidazioni, violazioni e minacce.

La particolarità di queste disposizioni è la platea dei suoi destinatari: il perimetro geografico è lo spazio stesso della piattaforma, motivo per cui gli Standard si applicano «a tutti, in tutto il mondo e a tutti i tipi di contenuti».

Gli Standard della community sono divisi in sei capitoli:

1. Violenza e comportamenti criminali;
2. Sicurezza;
3. Contenuti deplorevoli;
4. Integrità e autenticità;
5. Rispetto della proprietà intellettuale;
6. Richieste e decisioni relative ai contenuti.

Queste macro-categorie contengono al loro interno casistiche più specifiche. Ciascuna di esse (Tab. 1) viene presentata a partire da un fondamento che ne descrive gli obiettivi, seguito da esempi di contenuti non consentiti e

parola, o della stampa; o il diritto delle persone di riunirsi pacificamente in assemblea e di fare petizioni al governo per la riparazione dei torti».

contenuti che necessitano di integrazioni o schermate che ne impediscano la visione ai minori di 18 anni.

Tab. 1 *Standard della community*

Violenza e comportamenti criminiali	**Sicurezza**	**Contenuti deplorevoli**	**Integrità e autenticità**	**Rispetto della proprietà intellettuale**	**Richieste e decisioni relative ai contenuti**
Violenza e istigazione alla violenza	Suicidio e autolesionismo	Incitamento all'odio	Integrità dell'account e identità autentica	Proprietà intellettuale	Richieste degli utenti
Persone e organizzazioni pericolose	Sfruttamento sessuale, abusi e nudi di minori	Immagini forti e violente	Spam		Protezione aggiuntiva dei minori
Organizzazione di atti di violenza e promozione della criminalità	Sfruttamento sessuale di adulti	Immagini di nudo e atti sessuali di adulti	Sicurezza informatica		
Prodotti e servizi con restrizioni	Bullismo e intimidazioni	Adescamento	Comportamento non autentico		
Frode e raggiro	Sfruttamento di esseri umani		Disinformazione		
	Violazioni della privacy		Account commemorativi		

Fonte: elaborazione dell'autrice.

Soltanto nel 2019, Meta è intervenuta su 1873 miliardi di contenuti: 1800 miliardi sono stati classificati come *spam*, mentre 73 milioni di post rientravano in almeno una fattispecie contemplata negli Standard della community (Douek 2019, 11). Tra il 2018 e il 2022 le singole voci dei capitoli sono state modificate complessivamente 225 volte (Tab. 2).

Tab. 2 *Variazioni agli standard della community dal 2018 al 2022*

Standard della community	**Modifiche (2018-2022)**	**% sul totale**
Violenza e comportamenti criminali	77	36,84%
Sicurezza	55	26,32%
Contenuti deplorevoli	60	28,71%
Integrità e autenticità	15	7,18%

(continua)

(segue)

Rispetto della proprietà intellettuale	0	0,00%
Richieste e decisioni relative ai contenuti	2	0,96%
	209	100%

Fonte: elaborazione dell'autrice dal Registro modifiche aggiornate al marzo 2022.

Le principali revisioni (superiori al 5% del totale) hanno riguardato i campi (Fig. 1): violenza e istigazione alla violenza; prodotti e servizi con restrizioni; persone e organizzazioni pericolose[2]; incitamento all'odio; sfruttamento sessuale di adulti; integrità dell'account e identità autentica.

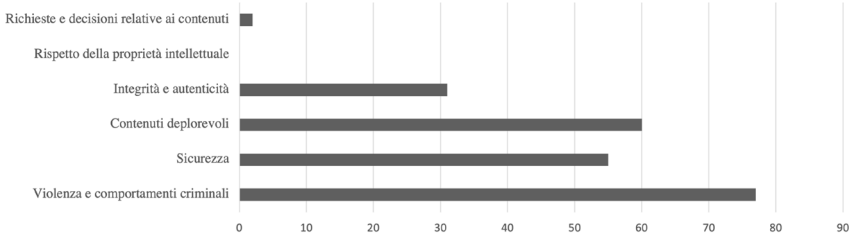

Fig. 1 Distribuzione delle modifiche apportate ai Capitoli degli «Standard della Community di Facebook dal 2018 al 2022».

L'amministrazione ha il potere di regolamentare le tecniche di moderazione che incideranno sulla quantità e sulla qualità dei prodotti immateriali generati dagli utenti che circolano sul proprio sito, definendo al contempo le sanzioni rivolte ai trasgressori. Nel caso in analisi, la moderazione del contenuto può avvenire: *ex ante*, prima che sia pubblicato, o *ex post*, a partire dalla sua pubblicazione; in maniera «reattiva», in base alle segnalazioni (*flag*), o «proattiva», mediante la ricerca di contenuti; in modalità «automatica, usufruendo di software, o «manuale», dal lavoro di segnalazione svolto dai revisori[3].

[2] Il documento è interno all'azienda e risale a circa nove anni fa. Dal 2021, Intercept sta svolgendo un lavoro di inchiesta sull'inserimento di soggetti e movimenti all'interno della lista che dimostrerebbe il tentativo di rinforzare le visioni politiche Usa a livello globale, punendo in maniera sproporzionata alcune comunità e soltanto in determinati contesti. Si veda Biddle (2021).

[3] Come osserva Monika Bickert, Vice presidente del Global policy management: «Una sfida è identificare le potenziali violazioni dei nostri standard in modo da poterle rivedere. La tecnologia può aiutare in questo senso. Utilizziamo una combinazione di intelligenza artificiale e di segnalazioni da parte delle persone per identificare post, immagini o altri contenuti che probabilmente violano i nostri standard comunitari».

Dunque, l'identificazione di un contenuto non idoneo alla permanenza segue principalmente due orientamenti: orizzontale, tra utenti; verticale, con l'intervento dell'Ia e dei moderatori di contenuti. Queste figure dotate del potere di sorveglianza e censura rappresentano l'ultimo filtro rimasto tra gli algoritmi e gli utenti (Franchi 2021) e l'ennesimo caso di lavoro precario prodotto dalla *gig economy*. Occorre distinguere tra l'azione volontaria delle persone (Crawford *et al.* 2016) e le mansioni ordinarie dei dipendenti della piattaforma, laddove queste operazioni vengono svolte con una capacità di monitoraggio e regolazione in tempo reale di cui i governi non dispongono. Lo stesso Zuckerberg ha ammesso che l'integrazione di sistemi di Ia alla forza lavoro dei moderatori avrebbe ridotto gli errori, ma non sarebbe mai diventata impeccabile (Douek 2019).

A ciascun dipendente spetta una formazione specifica comprendente tre fasi: pre-formazione (reperimento fonti), formazione pratica (minimo 80 ore affiancati a un istruttore), consulenza continua (documentazione aggiornata e kit di approfondimento).

Da un punto di vista tecnologico, sono stati fatti passi avanti nell'utilizzo di sistemi di Ia in questo settore (Bloch-Wehba 2020): con la *hash technology* i sistemi confrontano il contenuto pubblicato con un database di azioni ritenute illegali, soprattutto nel campo della pedopornografia, della protezione della libertà intellettuale e del contrasto alle forme di estremismo organizzato; con il *geo-blocking* si interviene sulla localizzazione dell'indirizzo Ip, che può essere associato a confini geografici per discriminare i contenuti sulla base delle norme vigenti in quello Stato; in materia di *spam*, i contenuti possono essere identificati dagli algoritmi in diversi modi basandosi sul comportamento degli utenti in relazione alla quantità di volte in cui un'informazione viene pubblicata o al suo target di destinazione (Klonick 2020, 2429-2431). La tecnologia prende provvedimenti su un nuovo contenuto quando rileva corrispondenze o somiglianze forti con un altro post che è già stato contrassegnato come non conforme. Meta ha provveduto a ingenti investimenti in questo campo:

Linformer e Reinforced integrity optimizer rintracciano l'*hate speech* in diverse aree geografiche;

SimSearchNet monitora la corrispondenza tra immagini, aiutando nella rilevazione di differenze sottili nei contenuti che contribuirebbero alla diffusione di disinformazione online (in particolare per temi legati a Covid-19);

Strumenti linguistici come *Xlm e Xlm-R*, che aiutano a interagire con contesti differenti, sono stati integrati con *RoBERTa*, per il *deep learning*.

Nonostante gli sforzi tecnologici, soltanto nei primi tre mesi del 2019 l'azienda ha ricevuto da parte degli utenti 25 milioni di richieste di appello (circa 275.000 al giorno) per le sue decisioni in materia di moderazione dei

contenuti. Tale circostanza rendeva evidente sia l'insufficienza dell'approccio industriale sia l'importanza di garantire coerenza e, quindi, giustizia (Douek 2019). Uno degli ambiti più complessi da valutare, per esempio, è la libertà di espressione perché corre il rischio di essere associata a una violazione per bullismo o *hate speech* mediante un mero controllo proattivo da parte delle macchine, incapaci di riconoscere le sfumature legate al contesto, alla lingua e alle norme sociali (*ibidem*). Alcune questioni richiedono un controllo caso per caso, come avviene per le notizie che, rappresentando un pubblico interesse, beneficiano di una eccezione alla norma. Tuttavia, restano un mistero per l'utente sia l'esistenza della coda di revisione, la mole di contenuti da revisionare, che i criteri di urgenza stabiliti da ciascun supervisore (Franchi 2021, 189) perché è praticamente impossibile stabilire un canale di comunicazione e di rassicurazione sulla presa in carico.

Su suggerimento di Noah Feldman (docente presso *Harvard law school*), e dopo due anni di consultazioni, nel 2020 Mark Zuckerberg ha introdotto l'Oversight board, Comitato per il controllo indipendente che ha il compito di esaminare i ricorsi e dirimere le controversie tra gli utenti e l'azienda, fornendo una spiegazione al mantenimento o meno del contenuto sulla piattaforma. Tra il 2020 e il 2022 ha selezionato 29 casi ed emanato 26 decisioni[4], esprimendosi su questioni che rispecchiano in proporzione i capitoli più problematici: contenuti deplorevoli (15), violenza e comportamenti criminali (10), sicurezza (1).

Con lo scoppiare della guerra in Ucraina è stato notato (Biddle 2021) come Meta avesse cambiato le sue politiche nei confronti del battaglione Azov in modo tale da mantenere i commenti a supporto, a patto che non inneggino al nazismo. L'azienda ha informato il Mob che avrebbe revocato la richiesta di indicazioni sulle normative riguardo a problemi di moderazione dei contenuti concernenti il conflitto tra Russia e Ucraina. A seguire, il Comitato per il controllo ha reso noto il suo parere, dichiarandosi deluso dalla decisione dell'azienda perché «il contributo dell'azienda nel difendere la libertà di espressione e i diritti umani è diventato sempre più importante».

3. La procedura di ricorso

Meta permette agli utenti di avviare una procedura di ricorso all'Oversight board (Mob) per contestare le decisioni dell'azienda in materia di contenuti pubblicati sulle piattaforme Facebook e Instagram in caso di disaccordo.

[4] Il dato è aggiornato a luglio 2022.

Sulla base delle segnalazioni ricevute, i membri del Comitato per il controllo selezionano i casi che necessitano di una revisione approfondita: spetta all'organismo, quindi, la definizione dei criteri di ammissibilità delle richieste per stabilire quali sono i livelli di gravità e priorità necessari per la riapertura di quei casi «idonei difficili, significativi e pertinenti a livello globale, in grado di arricchire le normative future», sulla base delle indicazioni contenute nello Statuto e, in particolare, in materia di libertà di espressione, autenticità, sicurezza, privacy, dignità. Fin dal primo momento viene chiarito che il Comitato lavorerà su un numero esiguo di ricorsi.

Per poter inviare un ricorso è necessario rispettare i seguenti requisiti: 1) la contestazione deve provenire da un account attivo (il proprietario deve essere in grado di accedervi); 2) l'appello è consentito soltanto qualora l'utente abbia già richiesto una revisione all'azienda; 3) le decisioni devono essere idonee al ricorso (a cui, se del caso, verrà assegnato un Id, cioè un numero di riferimento) per garantire efficacia nella tutela e rispetto delle leggi per ciascun paese; 4) la procedura deve essere avviata entro i 15 giorni successivi all'aggiornamento della decisione da parte di Facebook o Instagram.

Successivamente, a ciascun caso sarà assegnata una giuria a cui verranno trasmesse le informazioni utili alla comprensione del contesto – da parte dell'utente e da parte dell'azienda – per procedere al controllo «in conformità con le limitazioni sulla privacy e sulle leggi applicabili» (*ibidem*). La giuria potrà avvalersi di una consulenza da parte di esperti del settore per deliberare una decisione vincolante nei successivi 90 giorni, termine entro il quale l'azienda deve predisporne l'implementazione. Questo sottoinsieme del Comitato, comprendente almeno un membro proveniente dalla regione coinvolta, emanerà una bozza di decisione sulla quale tutti i membri potranno esprimersi prima che diventi definitiva.

Infine, il Comitato produrrà una spiegazione articolata e disponibile online, notificando al ricorrente la pubblicazione e concordando con le persone coinvolte quali informazioni rendere note al fine di autorizzare o impedire la riconoscibilità dei soggetti coinvolti. Tale autorizzazione, a valere sulla privacy dei singoli, può essere revocata in qualsiasi momento accedendo alla sezione del sito «Stato del caso». All'interno della dichiarazione potrà essere inserita una raccomandazione sulle normative rivolta all'azienda. Nei documenti che descrivono le fasi della procedura, in più di un'occasione si rammenta che Meta implementerà la decisione «purché tale applicazione non violi la legge».

4. Chi fa parte dell'Oversight board

Il Comitato per il controllo è un organismo indipendente e le sue attività, così come il finanziamento dei membri[5], sono supervisionate dai Garanti (*Trustee*): cinque figure trasversali, con competenze in ambito giuridico applicate alla tutela della libertà di espressione, esercitate in accademie, organizzazioni pubbliche e private.

Fungono da riferimenti normativi per la sua fondazione gli Statuti, l'Accordo Srl (tra i *Trustee* e società a responsabilità limitata), il Contratto fiduciario (tra Facebook e i *Trustee*), i Contratti dei membri (accordo con Facebook o una Srl), il Codice di condotta e il Contratto di servizio tra Facebook e una Srl (accordo per la fornitura di servizi). La disciplina del Mob è sancita nel suo Atto costitutivo composto da un'introduzione e sette articoli, dove vengono definiti: 1) l'autorevolezza; 2) l'ambito di competenza; 3) le procedure.

All'interno dell'Introduzione vengono presentati i valori fondanti dell'azienda e lo scopo del Comitato, ovvero «proteggere la libertà di espressione prendendo decisioni autonome e basate su principi determinati riguardo contenuti importanti ed emettendo pareri consultivi relativi alle normative sui contenuti di Facebook», operando in maniera trasparente nei confronti del pubblico e creando un rapporto di fiducia indipendente e irrevocabile con i *trustee*. I membri (articolo 1) sono pubblici ed esercitano un giudizio neutrale, indipendente imparziale. La sua composizione oscilla tra gli 11 e i 40 membri, con variazioni determinate dai casi. Ciascun membro resta in carica per tre anni e può essere rinnovato per un massimo di tre volte, con pena di decadenza al sussistere di condizioni effettive o percepite di conflitto di interessi per le quali è fatto esplicito divieto.

La composizione attuale del Mob comprende 23 membri, selezionati per aver maturato esperienze pregresse in ambiti come il diritto internazionale e le sue declinazioni, libertà di espressione e giornalismo, studi sui media e diritti digitali, sicurezza online, scienze politiche e diritti umani. Seppur la maggior parte provenga da contesti accademici, figurano tra le sue fila anche il premio Nobel per la pace Tawakkol Karman, l'ex Prima ministra danese Helle Thorning-Schmidt, lo scrittore Khaled Mansour e l'Ad di Pen America Suzanne Nossel. Il Comitato presenta un bilanciamento di genere e, per quanto concerne la distribuzione geografica, risultano numerosi gli statunitensi (6 su 23) mentre sono del tutto assenti personalità provenienti da contesti problematici per i rapporti con la piattaforma. Russia e Cina sono gli unici due Stati membri

[5] La retribuzione è emessa secondo un programma basato sull'adempimento dei compiti e non è influenzata o trattenuta sulla base del risultato delle decisioni del Comitato (si veda l'Atto costitutivo).

permanenti del Consiglio di sicurezza dell'Onu a non farne parte, sebbene vi sia una rappresentanza da Taiwan.

Il Comitato esercita i poteri collettivi (art. 1.IV) di: 1) richiedere a Facebook di fornire informazioni in modo tempestivo e trasparente; 2) interpretare gli Standard della community e altre normative rilevanti sui contenuti alla luce dei valori di Facebook; 3) indicare a Facebook di consentire o rimuovere contenuti; 4) indicare a Facebook di confermare o annullare un atto esecutivo; 5) emettere rapidamente spiegazioni scritte sulle sue decisioni.

È prevista l'elezione indiretta di co-presidenti che fungono da collegamenti con l'amministrazione del Comitato, i comitati direttivi, e dispongono di competenze in materia di gestione per selezionare i casi e proporre ulteriori candidati ai *trustee* (art. 1.VII); quest'ultima competenza non è esclusiva perché anche l'azienda e le persone possono avviare un procedimento di raccomandazione.

La richiesta di analisi dei contenuti può essere presentata sia dagli utenti che da Facebook (art. 2). Il Mob prenderà in esame la richiesta, salvo i casi in cui la decisione possa comportare responsabilità penale o sanzioni regolamentari. Ciascuna decisione avrà valore di precedente, motivo per cui le operazioni si svolgeranno con l'intento di raggiungere l'unanimità (art. 3.IV) ove possibile. Vengono, altresì, previste procedure speciali quali: «Riesame del comitato», qualora sia necessaria un'ulteriore analisi di una decisione del gruppo; «Analisi rapida», in circostanze eccezionali, compreso quando i contenuti possono avere ripercussioni gravi nel mondo reale, su segnalazione dell'azienda; «Richiesta di assistenza sulle normative», indipendentemente dai casi in sospeso e di tipo consultivo.

La risoluzione è vincolante (art. 4) e può fungere da precedente per le scelte future di Facebook nel dirimere controversie analoghe.

All'articolo 5 viene descritta la governance dell'Oversight board composta da:

- Comitato, a cui spetta il dovere di esaminare i casi, emanare decisioni, proporre nuovi membri (art. 5.I).
- *Trustee*, responsabile della gestione e dell'adesione del Comitato ai valori dello scopo dichiarato. Spettano ai garanti i doveri fiduciari conferiti da Facebook (art. 5.II).
- Facebook, che si impegna nella supervisione indipendente del Comitato nell'ambito delle decisioni sui contenuti e della conseguente attuazione. L'azienda – si legge – «supporterà il Comitato nella misura in cui le richieste siano tecnicamente e operativamente fattibili e coerenti con una ripartizione ragionevole delle risorse» (art. 5.III).

Infine, è possibile provvedere a una modifica parziale o sostanziale dello Statuto soltanto qualora si raggiunga l'approvazione della maggioranza dei *Trustee*, l'accordo di Facebook e la maggioranza dei membri del Comitato (art. 6). Quest'ultimo, tuttavia, «non pretende di applicare la legge locale» (art. 7).

Considerare il Comitato al pari di una Corte suprema con pubbliche funzioni, come più volte accaduto, produce una metafora costituzionale (Cowls *et al.* 2022, 2-3) che lo legittima nel contesto della governance delle piattaforme agita da attori privati. Questa associazione comporta serie implicazioni per la democrazia e per lo Stato di diritto, laddove si considerano pacifici concetti come il costituzionalismo, il senso della legge, l'autorità e il consenso pubblico senza il dovuto intervento da parte delle istituzioni esistenti. Come scriveva il professor Feldman (2018), si tratta di un sistema *quasi-legale* in un *quasi-stato* e definirlo tale ha creato un precedente. Una scelta non casuale che distingue accuratamente questo organismo da un ufficio di gestione dei rapporti con i clienti, che riconosce un elemento di pubblicità alla disputa tra le parti (Douek 2019). L'utilizzo nei documenti di lavoro e nelle pubblicazioni ufficiali dell'espressione «Corte suprema» ha contribuito alla formazione di un tacito assenso tendente a normalizzare l'esistenza di un meccanismo di giustizia che possa competere, se non sostituire, il potere razionale-legale. In questo senso, Mark Zuckerberg è il fondatore del Meta-diritto, che getta le basi per un bilanciamento nell'esercizio dei poteri di cui nessun cittadino l'ha investito, in una porzione di spazio che esiste nel perimetro determinato dalla sua volontà.

Pensare che il Mob sia il giudice ultimo in materia di nome sulla libertà di espressione a livello globale sarebbe un grave errore. Bisognerebbe, invece, considerare il suo ruolo nel contesto di un sistema legale domestico (Douek 2019, 7) a cui comunque va riconosciuto il merito di aver rimesso in moto il processo legislativo nella formulazione degli Standard della community, contribuendo, altresì, al coinvolgimento dei cittadini mediante un forum indipendente e umano di confronto sulle decisioni controverse.

5. Controllo e autonomia in Meta

Le tecnologie di Meta riescono a individuare, verificare e rimuovere la maggior parte dei contenuti ancor prima che questi vengano segnalati, incrociando le informazioni con le norme presenti negli Standard della community di Facebook e le Linee guida della community di Instagram, mediante un controllo che avviene su base settimanale. Nella maggior parte dei casi, la procedura (definita semplice) prevede la mera conferma della violazione. Quando questa mansione veniva svolta prevalentemente dal controllo umano, a partire dal-

le segnalazioni delle persone, i team trascorrevano la maggior parte del tempo controllando contenuti poco gravi o non in violazione. L'azienda ha dichiarato che tali decisioni «non erano utili a migliorare le nostre tecnologie per l'applicazione». L'approccio attuale tende a conferire priorità di un controllo umano a contenuti considerati potenzialmente pericolosi, «ossia su quelle decisioni per cui l'intervento umano è migliore rispetto a quello della tecnologia», al fine di preparare l'Ia a eseguire o non eseguire automaticamente un'azione.

Tale priorità sul caso da prendere in esame viene determinata da strumenti tecnologici e assegnata seguendo tre indicatori:

- Gravità. Quanto è probabile che il contenuto provochi un pericolo, sia online che offline?
- Diffusione. Con quale velocità è stato condiviso il contenuto?
- Probabilità di violazione. Quanto è probabile che il contenuto in questione violi effettivamente le nostre normative?

Rientrano in questo contesto i casi in cui l'intenzione del post non è chiara, la lingua è complessa e le immagini sono legate a situazioni contingenti. Grazie a questo lavoro cognitivo, che mobilita conoscenze e consapevolezze nei settori più disparati utili al raggiungimento di una decisione «sottile e spesso difficile», si allenano gli algoritmi a emulare la capacità umana: «Ogni volta che i controllori prendono una decisione, usiamo questa informazione per allenare la nostra tecnologia. Nel tempo, grazie a milioni di decisioni, la nostra tecnologia migliora, consentendoci di rimuovere un numero sempre maggiore di contenuti in violazione».

L'azienda ha provveduto alla creazione del team «Responsible Ai», composto da esperti di etica, ricercatori in ambito ingegneristico, politico, sociale e legale, chiamati a sviluppare linee guida, strumenti e processi per gestire le problematiche legate all'Ia, alle sue responsabilità e alla diffusione interna di queste risorse. A partire da questi studi sull'integrità, le squadre addette creano modelli di *machine learning* per portare a termine attività come l'analisi dei soggetti (persone) in foto o la comprensione del testo (contenuti). Queste previsioni, infine, aiutano nell'applicazione delle normative, stabilendo se è opportuno prendere provvedimenti punitivi (eliminazione o penalizzazione) o se è necessario sottoporre il caso all'attenzione dei revisori umani per ulteriori verifiche. Ad oggi questa tecnologia rileva e rimuove proattivamente più del 90% dei contenuti e account prima che questi vengano segnalati. L'azienda è al lavoro per evitare che le persone la aggirino.

Dal 2013 si applica il «controllo incrociato» per identificare i contenuti considerati come falsi positivi nel caso di pagine o gruppi influenti, cioè in grado di pubblicare contenuti potenzialmente virali e con rapida diffusione, per i quali si considera l'appartenenza (candidati politici, giornalisti, partner

commerciali, organizzazioni per i diritti umani), il numero di follower e l'ambito di competenza dell'entità. Il controllo in questione è formato da: *General secondary review*, *Early response e Secondary review*. Il Mob ha inserito questa fattispecie all'interno della sua relazione pubblicata nell'ottobre 2021, in cui si segnala che la risposta dell'azienda sul tema non è stata sufficientemente esaustiva. L'azienda ha chiesto al Comitato di fornire raccomandazioni su come: assicurare giustizia e oggettività, considerando il contesto; condurre i controlli incrociati e promuovere la trasparenza; definire i criteri per determinare chi includere nel rispetto dell'imparzialità.

Esistono almeno due aree su cui il Comitato per il controllo dovrebbe esprimersi (Douek 2019, 40-41): la prima fa riferimento al *ranking* algoritmico e a come influenza la circolazione (e la visibilità) dei contenuti; la seconda all'applicazione delle politiche pubblicitarie. Entrambe rappresentano un elemento chiave per la strategia aziendale di Meta, che interviene sul *design* del prodotto e sul profitto, oltre che questioni rilevanti in materia di moderazione, soprattutto se comportano una minore esposizione e distribuzione del post o della pagina presa in esame. Questa eventualità, meno trasparente rispetto alla rimozione del contenuto, non viene segnalata all'utente tramite una notifica.

Pur disponendo di un numero esiguo di casi su cui il Mob si è espresso, a fronte di circa un milione di ricorsi avanzati dagli utenti di Facebook e Instagram nel biennio, è possibile avanzare alcune riflessioni qualitative. Il Mob ha condiviso con l'azienda 6 decisioni e ne ha revocate 19 (pari al 76% delle deliberazioni prese in esame)[6].

In particolare, il dibattito emerso in occasione della deliberazione 2020-004-Ig-ua contiene importanti elementi di analisi. Nell'ottobre 2020, un utente in Brasile ha pubblicato un'immagine su Instagram in occasione del *Pink october*, una campagna internazionale di sensibilizzazione sul cancro al seno. Il post è stato rimosso da un sistema automatizzato che aveva associato le foto alla violazione degli Standard della community riguardanti immagini di nudo e atti sessuali di adulti. Come si legge sul sito dell'Oversight board, «in seguito alla selezione di questo caso da parte del Comitato, Facebook ha ripristinato il post, stabilendo che si trattava di un errore». L'Oversight board ha precisato che la rimozione errata di un contenuto denota la mancanza di un adeguato controllo umano. I sistemi non sono stati in grado di riconoscere le parole chiave «cancro al seno» per poter distinguere l'effettivo scopo delle foto, interferendo con la libertà di espressione. Veniva chiarito, altresì, che il diverso trattamento riservato ai capezzoli maschili e femminili si traduce in

[6] Dati aggiornati a luglio 2022. Il caso 2020-001-Fb-ua non rientra nel computo perché l'utente ha rimosso il contenuto prima della deliberazione, interrompendo la procedura di controllo del Comitato.

un'automazione imprecisa dell'applicazione delle regole, che influenza in modo sproporzionato la libertà di espressione delle donne. L'ulteriore criticità evidenziata riguardava la motivazione dichiarata all'utente: la decisione sarebbe stata presa considerando le Linee guida della community di Instagram, a cui, però, vanno integrati gli Standard della community di Facebook, tra i quali era prevista l'eccezionalità della fattispecie in oggetto.

Infine, occorre considerare che in alcuni Paesi l'approvazione di leggi nazionali ha reso più complessa l'applicazione delle norme decise dall'azienda. In Austria il «Communication platform act» e in Germania il «Network enforcement act (NetzDg)» impongono una procedura *ad hoc* per i reclami su contenuti illeciti sui social network. L'approccio adottato considera tre ipotesi: se il reclamo incontra corrispondenza tra le Linee guida della community e quanto affermato nel testo di legge, il contenuto viene rimosso a livello globale e il processo di analisi di conclude; se il contenuto non viola le Linee guida della community, si analizza la legittimità in base alla segnalazione; se il contenuto viola il Codice penale nazionale, l'accesso viene disabilitato in quel paese specifico.

Il Mob ha revocato la decisione di rimuovere un post su Instagram per l'utilizzo discriminatorio di alcune parole che in arabo si riferiscono in chiave dispregiativa agli uomini con «effeminate mannerisms». L'errore di Meta è stato quello di non riconoscere il contenuto come eccezione alla normativa in materia di incitamento all'odio. Il caso 2022-003-Ig-ua è emblematico perché la piattaforma è diventata uno spazio di discussione per le narrazioni queer nella cultura araba. Un account pubblico aveva condiviso un carosello (un post contenente fino a 10 immagini con un'unica didascalia) per spiegare in arabo e in inglese come alcuni termini fossero utilizzati con lo scopo di offendere la comunità Lgbtqia+. Il contenuto è stato rimosso e ripristinato varie volte prima che venisse sottoposto al controllo interno, ma il Board «teme che Meta non applichi in modo coerente le esenzioni alle affermazioni dei gruppi emarginati, previste dalla normativa» e, anzi, ne incentiva il coinvolgimento per elaborare elenchi di parole che tengano conto delle sfumature per ciascun «mercato». Tutti i moderatori coinvolti – che parlavano correntemente l'arabo – hanno confermato la violazione, dimostrando che le linee guida (in lingua inglese) e la formazione potrebbero non essere sufficienti.

6. Conclusioni

Ogni settimana vengono segnalati migliaia di contenuti da parte degli utenti. Spesso intercorrono valutazioni e opinioni personali, conflitti tra grup-

pi o volontà di arrecare danno ad altri utenti. Ciò che bisogna considerare è che molti di questi contenuti, in realtà, non violano gli Standard della community (Klonick 2020, 2432). Questo è il motivo per cui Meta chiede di compilare un modulo al momento della segnalazione: si richiede di selezionare il tipo di offesa e le motivazioni per cui il contenuto debba essere rimosso od oscurato. Tuttavia, alcune violazioni non sono presenti nel volano di opzioni selezionabili. In aggiunta, le questioni concernenti il controllo e l'autonomia su dati e contenuti non vengono condivise sufficientemente con chi decide di utilizzare la piattaforma. Come ha ammesso l'Oversight board nell'ottobre 2021: «Un tema è emerso chiaramente: Facebook non mostra chiarezza nei confronti delle persone che usano le sue piattaforme. Abbiamo costantemente visto utenti che dovevano tirare a indovinare il motivo per cui Facebook avesse rimosso i loro contenuti». Alcuni casi destano maggiori preoccupazioni perché celano la possibilità che l'utente – cancellando il contenuto prima di ricevere un riscontro, vedendosi impossibilitato a sostenere la qualità delle proprie ragioni – rinunci a contestare una violazione della sua libertà o, ancor di più, a esprimersi.

L'analisi del ruolo del Comitato per il controllo rinvia a una dicotomia tra diritto pubblico e diritto privato determinata dal rapporto ambivalente tra cittadino e azienda, dove il primo agisce in qualità di cliente, avendo sottoscritto un contratto al momento dell'iscrizione con l'accettazione dei termini e delle condizioni di utilizzo del prodotto offerto dalla piattaforma. Quest'ultima può migliorare grazie alle segnalazioni e ai ricorsi degli utenti, il che sposta ulteriormente il piano della narrazione: il cliente dovrà fidarsi della volontà dell'azienda di agire nel rispetto dei Diritti umani. Il risarcimento previsto è la reintegrazione di un contenuto oscurato che sostanzia l'impegno a tutela della libertà di espressione.

È giusto chiedersi se l'ordine del giorno del Comitato segua l'agenda politica del momento? Indirettamente questo accade ogni qualvolta l'azienda decide di modificare repentinamente gli Standard della community per adattarli a un bisogno contingente. A ciascuna modifica corrisponde la conseguente violazione delle norme che ridisegnano l'esperienza dell'utente sulla piattaforma, che vive relazionandosi con diverse aree geografiche, nazioni e culture e considerando ciascuna un «mercato». Risulta intuitivo immaginare che i conflitti possano riorientare il *business plan*. Anche in caso di errore, infine, gli addetti al controllo non sono tenuti a giustificare le ragioni dele loro decisioni. Nonostante ciò, il Mob riconosce all'azienda responsabilità di attuazione e tutela dei diritti umani e ritiene che «l'eccessiva moderazione delle opinioni degli utenti appartenenti a gruppi minoritari perseguitati» rappresenti una minaccia grave e diffusa alla loro libertà di espressione perché queste comunità che ne sopportano il peso «sono il riflesso di scelte di progettazione dei sistemi di applica-

zione». Secondo Franchi (2021, 207), il compito dei revisori non è la difesa delle persone, bensì la difesa della macchina dal sospetto che si possa metterne in dubbio l'infallibilità e per scongiurare la richiesta di una sua disattivazione temporanea o definitiva.

L'articolo esplora peculiarità e ambiguità di un organismo creato *ad hoc* e a cui viene richiesto di svolgere un servizio di tutela dei diritti fondamentali, muovendo, tuttavia, nel campo scelto dall'azienda, che si potrebbe definire di natura commerciale. I margini di discrezionalità sul compimento delle azioni del Mob lascia all'azienda un potere di esecuzione non ignorabile.

Emerge, infine, una ulteriore discrasia nelle dichiarazioni dell'azienda determinata dall'oscillazione tra l'aiuto offerto e percepito che i revisori e l'Ia si scambiano vicendevolmente. Da un lato, il controllo umano aiuta gli algoritmi ad allenarsi; dall'altro, tale controllo rischia di introdurre una percentuale di errore nel funzionamento ordinario. L'Ia, a sua volta, rileva e rimuove i contenuti in autonomia ma, secondo il Comitato, la maggior parte delle volte sbaglia.

Riferimenti bibliografici

Biddle, S. (2021), *Revealed: Facebook's Secret Blacklist of «Dangerous Individuals and Organizations»*, «The Intercept», 12 ottobre, https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/.

Bloch-Wehba, H. (2020), *Automation in Moderation*, in «Cornell International Law Journal», 53, pp. 41-96.

Cowls, J., Darius, P. Santistevan, D. e Schramm, M. (2022), *Constitutional Metaphors: Facebook's «Supreme Court« and the Legitimation of Platform Governance*, in «Ny: Social Science Research Network», doi:10.2139/ssrn.4036504.

Crawford, K. e Gillespie, T. (2016), *What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint*, «New Media & Society», 18 (3), pp. 410-428.

Douek, E. (2019), *Facebook's Oversight Board. Move Fast With Stable Infrastructure and Humility*, in «North Carolina Journal of Law and Technology», 21(1), pp. 1-79.

Feldamn, N. (2018), *A Supreme Court for Facebook. Global Feedback and Input on the Facebook Oversight Board*, Appendix D, https://about.fb.com/wp-content/uploads/2019/06/oversight-board-consultation-report-appendix.pdf. Consultato il 10 aprile 2022.

Franchi, J. (2021), *Gli obsoleti. Il lavoro impossibile dei moderatori di contenuti*, Milano, Agenzia X.

Klonick, K. (2020), *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, in «Yale Law Journal», 129, pp. 2418-2499.

Klonick, K. (2021), *Inside the Making of Facebook's Supreme Court*, The New Yorker, 12 febbraio, https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court.

Sitografia

https://about.fb.com/news/2018/04/comprehensive-community-standards/.
https://transparency.fb.com/it-it/enforcement/detecting-violations/training-review-teams.
https://transparency.fb.com/it-it/enforcement/detecting-violations/how-enforcement-technology-works/.
https://transparency.fb.com/it-it/enforcement/detecting-violations/investing-in-technology/.
https://transparency.fb.com/it-it/enforcement/detecting-violations/technology-detects-violations/.
https://transparency.fb.com/it-it/policies/improving/reviewing-enforcement-incidents/.
https://transparency.fb.com/it-it/policies/improving/prioritizing-content-review/.
https://transparency.fb.com/it-it/oversight/oversight-board-cases.
https://www.facebook.com/help/instagram/428536715033518/?helpref=related_articles.
https://www.facebook.com/help/instagram/1787585044668150/?helpref=related_articles.
https://transparency.fb.com/it-it/policies/community-standards/.
https://about.facebook.com/code-of-conduct/.
https://oversightboard.com/decision/IG-7THR3SI1/.
https://oversightboard.com/decision/FB-KBHZS8BL/.
https://www.oversightboard.com/news/215139350722703-oversight-board-demands-more-transparency-from-facebook/.
https://oversightboard.com/news/382264103827624-protecting-freedom-of-expression-and-human-rights-in-ukraine-and-russia/.
https://oversightboard.com/news/464999558726685-oversight-board-appoints-new-board-members/.
https://oversightboard.com/decision/IG-2PJ00L4T/.
https://oversightboard.com/governance/.
https://oversightboard.com/appeals-process/.